(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Publication number: **0 248 609**
A1

(12) EUROPEAN PATENT APPLICATION

(21) Application number: **87304793.0**

(51) Int. Cl.⁴: **G10L 7/08**

(22) Date of filing: **29.05.87**

(30) Priority: **02.06.86 GB 8613327**

(43) Date of publication of application:
09.12.87 Bulletin 87/50

(84) Designated Contracting States:
AT BE CH DE ES FR GB GR IT LI LU NL SE

(71) Applicant: **BRITISH TELECOMMUNICATIONS plc**
British Telecom Centre 81 Newgate Street
London EC1A 7AJ(GB)

(72) Inventor: **Forse, Nicholas John Arnold**
28 Netley Close Stoke Park
Ipswich Suffolk(GB)

(74) Representative: **Roberts, Simon Christopher et al**
BRITISH TELECOM Intellectual Property Unit
13th Floor 151, Gower Street
London, WC1E 6BA(GB)

(54) Speech processor.

(57) In a speech processor such as a speech recogniser, the problem of distortion of extracted features caused by adaption of the input automatic gain control (AGC) during feature extraction is solved by storing the AGC's gain coefficient along with the energy level of each extracted feature. At the end of the sampling period the stored gain coefficients are set equal to the minimum stored coefficient and the associated energy levels adjusted accordingly. The AGC circuit may comprise a digitally switched attenuator under the control of a microprocessor performing the speech recognition.

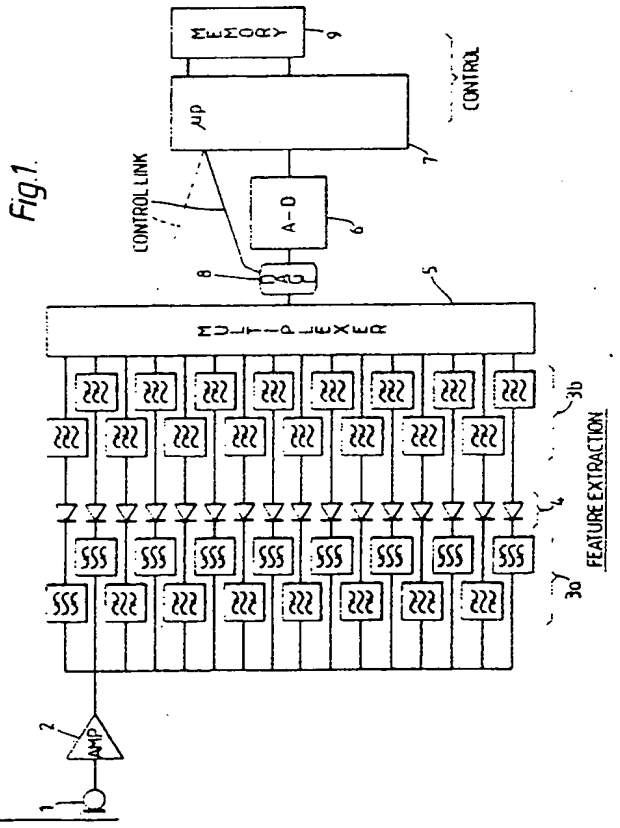


Fig.1.

EP 0 248 609 A1

SPEECH PROCESSOR

This invention relates to speech processors having automatic gain control, and in particular to speech recognisers.

Automatic speech recognisers work by comparing features extracted from audible speech signals. Features extracted from the speech to be recognised are compared with stored features extracted from a known utterance.

For accurate recognition it is important that the features extracted from the same word or sound then spoken at different times are sufficiently similar. However, the large dynamic range of speech makes this difficult to achieve, particularly in areas such as hands-free telephony where the sound level received by the microphone can vary over a wide range. In order to compensate for this speech level variation, most speech recognisers use some form of automatic gain control (AGC).

The AGC circuit controls the gain to ensure that the average signal level used by the feature extractor is as near constant as possible over a given time period. Hence quiet speech utterances are given greater gain than loud utterances. This form of AGC performs well when continuous speech is the input signal since after a period of time, the circuit gain will optimise the signal level to give consistent feature extraction. However, in the absence of speech, the gain of the AGC circuit will increase to a level determined by the background noise, so that at the onset of a speech utterance the gain of the AGC circuit will be set too high. During the utterance the gain of the circuit is automatically reduced, the speed of the gain change being determined by the 'attack' time of the AGC. The start of the utterance is thus subjected to a much greater gain and any features extracted will have a much greater energy content than similar features extracted later, when the gain has been reduced.

This distortion effect is dependent on the input signal level; the higher the speech level the larger is the distortion. Hence the first few features extracted will not correspond to the notionally similar stored features, and this can often result in poor recognition performance.

The present invention seeks to provide a solution to this problem.

According to the present invention there is provided a speech processor comprising an input to receive speech signals; signal processing means to extract spectral parameters from said speech signals; an analogue to digital converter to digitise said extracted parameters; an automatic gain control means to control the signal level applied to said converter; characterised in that the spectral parameters are stored at least temporarily, and for each such stored parameter a gain coefficient indicative of the gain applied by the gain control means is also stored; and in that at the end of a sampling period the gain coefficients stored in that period are, if different, set equal to the lowest gain coefficient stored in that period, the magnitudes of the corresponding stored spectral parameters being adjusted proportionally.

In a speech processor according to the invention, configured as a speech recogniser, automatic gain control is provided by a digitally switched attenuator, the gain of which is determined by the microprocessor performing the speech recognition. The microprocessor controls the gain to ensure that the dynamic range of the Analogue to Digital converter (which occurs between feature extraction and the microprocessor controlling the recogniser even when analogue AGCs are used) is not exceeded (except during the adaption of the AGC). The principal difference between the known analogue AGCs and the system according to the invention is that in the latter the microprocessor has control of the gain setting and can therefore store the gain used for each feature extracted. After the utterance has finished, the microprocessor can determine the optimum gain setting for the complete utterance. All the features stored are then normalised to this optimum gain setting. By this means a consistent set of features are extracted independent of the input signal gain.

Embodiments of the invention will be further described and explained by the reference to the accompanying drawing, in which

Figure 1 is a schematic diagram of a speech recogniser according to the present invention.

Throughout this patent application the invention is described with reference to a speech recogniser utilising template-matching, but as those skilled in the art will be aware, the invention is equally applicable to any of the conventional types of speech recogniser, including those using stochastic modelling, Markov chains, dynamic-timewarping and phoneme-recognition.

Speech recognition is based on comparing energy contours from a number (generally 8 to 16) of filter channels. While speech is present, the energy spectrum from each filter channel is digitized with an Analogue to Digital (A-D) converter to produce a template which is stored in a memory.

The initial stage of recognition is known as 'training' and consists of producing the reference templates by speaking to the recogniser the words which are to be recognised. Once reference templates have been made for the words to be recognised, recognition of speech can be attempted.

When the recogniser is exposed to an utterance, it produces a test template which can be compared with the reference templates in the memory to find the closest match.

The fundamental elements of the speech recogniser according to the present invention are shown in Figure 1. Voice signals received by the microphone 1 and amplified by amplifier 2 are passed to a filter bank 3a. In the filter bank the voice signals are filtered into a plurality (in this case 16) of frequency bands, and the signals are rectified by rectifier 4. The filtered and rectified signals are smoothed by low pass filters 3b and then sequentially sampled by a multiplexer 5 which feeds the resultant single channel signal to the DAGC circuit 8 which is turn feeds an Analogue to Digital converter 6 from which the digitized signal stream is passed to the controlling microprocessor 7.

The multiplexer addresses each filter channel for 20 microseconds before addressing the next one. At the end of each 10 millisecond time slot, each channel's sampled energy for that period is stored. The templates, which are produced during training or recognition, consist of upto 100 timeslot samples for each filter channel.

The digital AGC operates in the following way. Each time the multiplexer addresses a filter channel, the microprocessor assesses the channel's energy level to determine whether the A-D converter has been overloaded and hence that the gain is too high. When the microprocessor determines that the gain is too high it decrements the AGC's gain by 1 step, which corresponds to a reduction in gain of 1.5dB, and looks again at the channel's energy level. The multiplexer does not cycle to the next channel until the microprocessor has determined that the gain has been reduced sufficiently to prevent overloading of the A-D converter. When the multiplexer does cycle to the next filter channel, the gain of the AGC circuit is held at the new low level unless that level results in the overloading of the A-D converter with the new channel's energy level, in which case the gain is incremented down as previously described. When the multiplexer has addressed the final filter channel, the microprocessor normalises the energy levels of all the channels by setting their gain coefficients (which have been stored together with the energy level information in memory associated with the microprocessor) to the new minimum established by the microprocessor. In this way a consistent set of features are extracted independent of the initial output signal gain and any changes in the gain during formation of the template.

The speech recogniser is also required to detect the beginning and end of the speech or word with a high degree of accuracy. The speech recogniser according to the present invention uses the following technique:

A. The energy level of the background noise is measured and stored for 32 time slots (at 10 milliseconds a sample) while simultaneously adjusting (reducing) the gains of the AGC circuit as described above to cope with the maximum noise energy.

B. The maximum energy sample is found by adding all the filter values for each time slot, dividing by 16 (the number of filter channels) and multiplying by a gain factor corresponding to the gain of the DAGC circuit, and then comparing each time slot to find the maximum.

C. The threshold which needs to be exceeded before speech is deemed to be present is set to be equal to 1.5 times the maximum noise energy determined in Step B.

D. The average noise energy for each filter channel is found and stored (for each channel it is the sum of energies over all 32 time slots, divided by 32) to establish a noise template.

E. Thereafter, the filter bank is scanned every 10 milliseconds and the data is stored in a temporary cyclic store, of 100 time samples, until the average filter energy exceeds the noise/speech threshold calculated in C.

F. If the noise/speech threshold is not exceeded after 32 samples, a check is performed to ensure that the gain of the DAGC circuit is not set too low. This is done by looking at the maximum filter channel value stored in those 32 time slots. If that maximum level is 1.5dB or more below the maximum acceptable input level for the A-D converter, the gain of the AGC is incremented by 1 to increase the gain by 1.5dB. If the threshold is not exceeded after 32 samples and the DAGC setting is correct, then the noise/speech threshold is recalculated by finding the maximum energy over the last 32 samples (as in B) and multiplying by 1.5 (as in C).

G. Once the noise/speech threshold has been exceeded the filter bank is scanned every 10 milliseconds and the filter data is stored in memory, to form the speech templates, until either 100 samples have been entered or until the energy level drops below the noise/speech threshold for 20 consecutive samples. As described above, if during the data input the A-D converter is overloaded, the AGC setting is decremented by 1 and the data for that filter channel is reprocessed. If during the scan

of the 16 filter channels the gain of the DAGC circuit is reduced, the data from all 16 channels is re-input so that all the filter data corresponds to the same AGC setting. The AGC value used is recorded in memory along with the filter data. The AGC setting used at the start of each time slot is taken from the previous time frame, hence the gain can only be reduced (not increased) during the speech processing phase. This is not a problem since at the end of the template period all the template data is normalised to a uniform AGC setting.

H. To ensure that the start of speech was not missed by the speech/noise detector threshold, the 15 time samples prior to speech detection are transferred from the temporary cyclic store to the front of the 'speech' template.

I. If more than 100 samples were processed prior to speech being detected, the noise template is recalculated by analysing (as in D) the oldest 32 time frames in the temporary cyclic store. If less than 100 samples were processed prior to speech being detected, the noise template established in step D is used in the following steps.

J. The minimum gain setting of the AGC over the speech template is then found and both the speech and noise templates are normalised to this setting, which results in both templates containing the values that would have been entered had that gain been used from the start.

K. The normalised noise template is then subtracted from every time frame of the normalised speech template.

L. The maximum energy in the normalised speech template is now found and a new noise/speech threshold calculated - equal to the maximum energy minus 18dB. This new threshold is used to scan the normalised speech template to determine the start and finish points of the speech.

M. The speech template is then truncated to the start and finish points and is either stored in memory (training) or is used for recognition. The following tabular example represents the values stored after measuring the background noise for 320 milliseconds (32 time slots of 10 milliseconds each).

Filter bank number.

	DAGC	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Real AV energy
	4	210	220	232	245	224	216	167	188	176	234	250	177	134	170	213	209	408
	4	210	218	230	250	220	222	170	190	173	230	255	170	137	172	215	212	409
	4	210	222	234	247	216	225	171	189	178	233	253	171	140	170	214	208	410
	4	213	220	231	251	218	223	166	184	174	230	250	168	133	165	220	216	408
	4	215	217	228	253	220	220	160	186	180	231	254	166	132	164	223	220	409
	4	210	220	232	245	224	216	172	187	177	235	253	160	130	172	214	207	407
	4	210	220	232	245	224	216	167	188	176	234	250	177	134	170	213	209	408
T	4	211	218	230	250	220	222	170	190	173	230	255	170	137	172	215	212	409
I	4	210	222	234	247	216	225	171	189	178	233	253	171	140	170	214	208	410
M	4	213	220	231	251	218	223	166	184	174	230	250	168	133	165	220	216	408
E	4	215	217	228	253	220	220	160	186	180	231	254	166	132	164	223	220	409
	4	210	220	232	245	224	216	172	187	177	235	253	160	130	172	214	207	407
F	4	210	220	232	245	224	216	167	188	176	234	250	177	134	170	213	209	408
R	4	211	218	230	250	220	222	170	190	173	230	255	170	137	172	215	212	409
A	4	210	222	234	247	216	225	171	189	178	233	253	171	140	170	214	208	410
M	4	213	220	231	251	218	223	166	184	174	230	250	168	133	165	220	216	408
E	4	215	217	228	253	220	220	160	186	180	231	254	166	132	164	223	220	409
S	4	210	220	232	245	224	216	172	187	177	235	253	160	130	172	214	207	407
	4	210	220	232	245	224	216	167	188	176	234	250	177	134	170	213	209	408
	4	211	218	230	250	220	222	170	190	173	230	255	170	137	172	215	212	409
	4	210	222	234	247	216	225	171	189	178	233	253	171	140	170	214	208	410
	4	213	220	231	251	218	223	166	184	174	230	250	168	133	165	220	216	408
	4	215	217	228	253	220	220	160	186	180	231	254	166	132	164	223	220	409
	4	210	220	232	245	224	216	172	187	177	235	253	160	130	172	214	207	407
	4	210	222	234	247	216	225	171	189	178	233	253	171	140	170	214	208	410
	4	213	220	231	251	218	223	166	184	174	230	250	168	133	165	220	216	408
	4	215	217	228	253	220	220	160	186	180	231	254	166	132	164	223	220	409
	4	210	220	232	245	224	216	172	187	177	235	253	160	130	172	214	207	407
	4	210	220	232	245	224	216	167	188	176	234	250	177	134	170	213	209	408
	4	211	218	230	250	220	222	170	190	173	230	255	170	137	172	215	212	409
	4	213	220	231	251	218	223	166	184	174	230	250	168	133	165	220	216	408
	4	215	217	228	253	220	220	160	186	180	231	254	166	132	164	223	220	409

Average noise template :-

212 219 231 248 220 220 167 187 176 232 252 169 134 169 217 212

A DAGC value of 4 is equivalent to a 6dB attenuation of the signal going into the A/D, hence to calculate the "real" energy all the filter bank values above would have to be doubled.

Maximum real energy (averaged over all filters) was:- 410

Threshold to be exceeded to start/end template recording:- 615

Because the invention's primary application is to voice recognition it has been described with reference to that application. However, as those skilled in the art will be aware, the invention is not only applicable to voice recognition, but is applicable to practically any situation where voice signals are processed for feature extraction.

The speech processor according to the present invention is particularly suitable for use in applications where background noise and variations in the level of that background noise are a problem for known speech processors. One such application is in hands-free telephony, and in particular hands-free telephony involving cellular radio terminals. Such terminals are frequently used in cars, where it is convenient to use speech recognition to provide hands-free call connection and dialling. The problem arises however that wind, road and engine noise fluctuate widely and make accurate recognition of speech difficult. Clearly, if speech recognition for hands-free telephony is to be fully acceptable in this application it is necessary that the recogniser accepts and acts correctly in response to voiced commands in the presence of background noise, without routinely requiring that the commands be repeated.

The improved accuracy of recognition provided by the present invention is of particular advantage in this application.

Claims

1. A speech processor comprising an input to receive speech signals; signal processing means to extract spectral parameters from said speech signals; an analogue to digital converter to digitise said extracted parameters; an automatic gain control means to control the signal level applied to said converter; characterised in that the spectral parameters are stored at least temporarily, and for each such stored parameter a gain coefficient indicative of the gain applied by the gain control means is also stored; and in that at the end of a sampling period the gain coefficients stored in that period are, if different, set equal to the lowest gain coefficient stored in that period, the magnitudes of the corresponding stored spectral parameters being adjusted proportionally.

2. A speech processor as claimed in claim 1 in which each extracted spectral parameter corresponds to the energy content of a particular frequency band in a time slot of length t, further characterised in that for each extracted parameter the signal level applied to the analogue to digital converter is determined in a small fraction of time t, and if the signal level is greater than a predetermined level the gain is reduced and the signal level re-assessed, the signal strength assessment and the gain reduction being repeated within time slot t until the signal level is at a finalised level not exceeding said predetermined level.

3. A speech processor as claimed in claim 2 wherein said predetermined level is equal to the maximum level which does not exceed the dynamic range of the analogue to digital converter.

4. A speech processor as claimed in claim 2 or claim 3 wherein in a single time slot of length t spectral parameters are established for a plurality of discrete frequency bands, further characterised in that the different frequency bands are addressed sequentially, with the finalised gain coefficient of any frequency band being used as the initial gain coefficient of the next addressed frequency band.

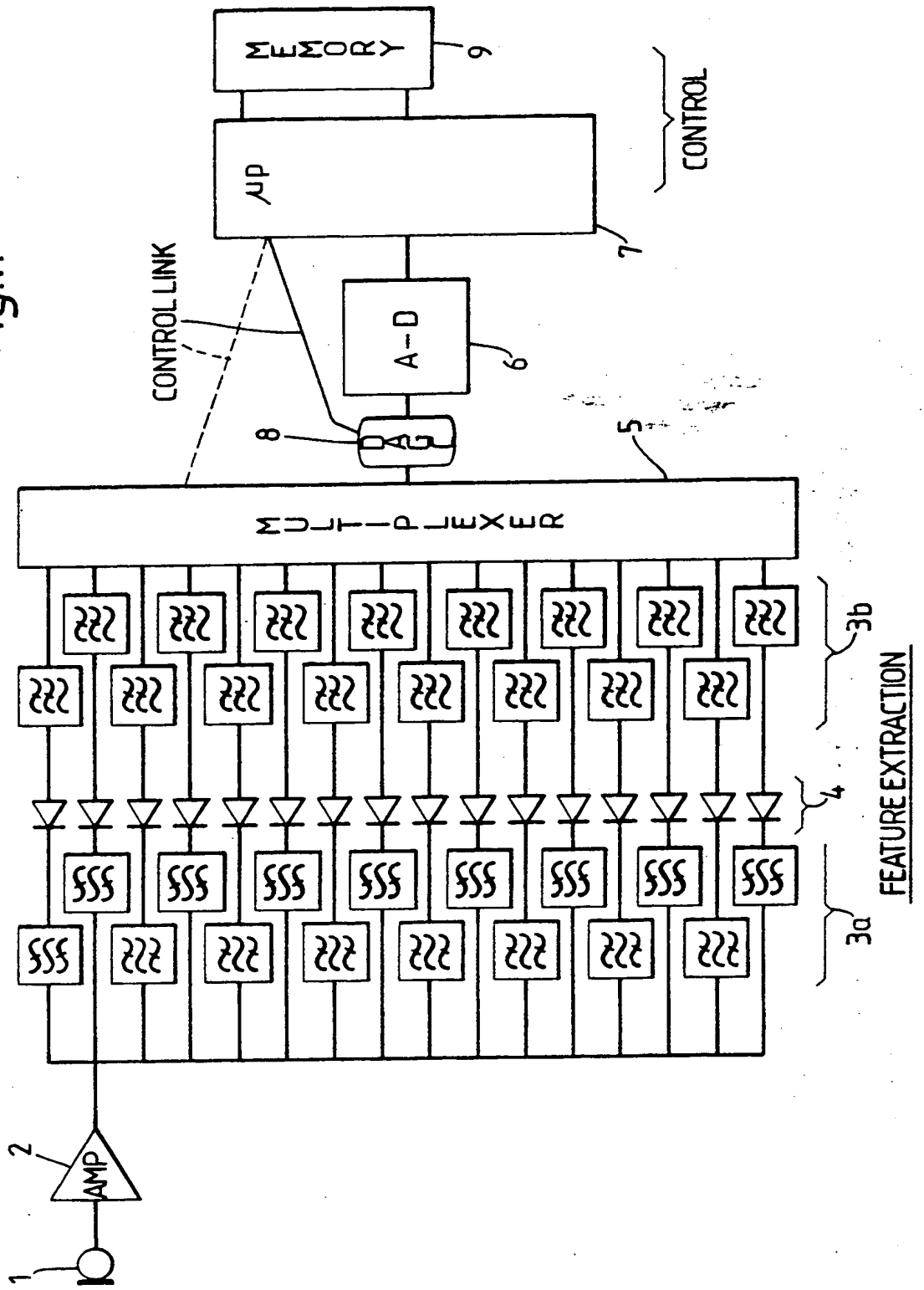
5. A speech processor as claimed in any one of claims 2 to 4 wherein the sampling period is made up of a plurality of time slots of length t.

6. A speech processor as claimed in any one of the preceding claims, configured as a speech recogniser.

7. A speech processor as claimed in any one of the preceding claims, wherein the gain control means comprises a digitally switched attenuator under the control of a microprocessor one of whose inputs is connected to the digitised output of the analogue to digital converter, the gain of the attenuator being determined by the microprocessor.

8. A cellular radio terminal comprising a speech recogniser for selecting functions in response to voiced instructions, characterised in that the speech recogniser comprises a speech processor as claimed in any one of claims 1 to 5.

Fig.1.



European Patent Office

EUROPEAN SEARCH REPORT

Application number

EP 87 30 4793

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int. Cl.4)
A	IEEE JOURNAL OF SOLID-STATE CIRCUITS; VOL: SC-19, no. 6, December 1984, pages 956-963, IEEE, New York, US; B. GILBERT: "A monolithic 16-channel analog array normalizer" * Abstract; figure 2 *	1,3,6	G 10 L 7/08
A	US-A-3 411 153 (R.W. STEELE) * Abstract *	1,3	
A	GB-A-2 107 102 (VERBEX CORP.) * Page 5, lines 52-55 *	1	
A	IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, 14th-16th April 1983, Boston, vol. 2, pages 511-514, IEEE, New York, US; J.A. FELDMAN et al.: "A custom IC for automatic gain control in LPC vocoders" * Introduction *	1-3,7	
A	IEEE TRANSACTIONS ON COMMUNICATIONS, vol. COM-30, no. 4, April 1982, pages 574-580, IEEE, New York, US; K. NIWA et al.: "A new channel bank with block companding" * Paragraph II,A: "Companding procedure" *	1-3	

-/-

TECHNICAL FIELDS SEARCHED (Int. Cl.4)

G 10 L 7/08

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	04-09-1987	ARMSPACH J.F.A.M.

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO Form 1503 03 82

European Patent Office

EUROPEAN SEARCH REPORT

Application number

EP 87 30 4793

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int. Cl.4)
A	IEEE TRANSACTIONS ON COMPUTERS, vol. C-20, no. 4, September 1971, pages 972-978, New York, US; L.C.W. POLS: "Real-time recognition of spoken words" * Page 975, right-hand column, lines 25-28 *	1	
A	IEEE SPECTRUM, vol. 8, no. 8, August 1971, pages 57-69, New York, US; G.L. CLAPPER: "Automatic word recognition" * Page 61, left-hand column, lines 35-39 *	1	

TECHNICAL FIELDS SEARCHED (Int. Cl.4)

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	04-09-1987	ARMSPACH J.F.A.M.

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO Form 1503 03 82